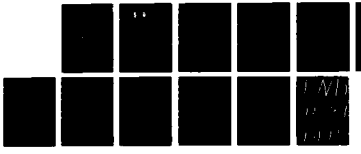
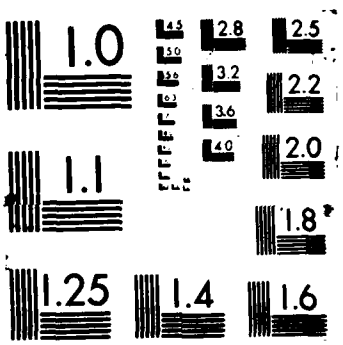
AD-A186 164 SOME PROPERTIES OF MAXIMUM LIKELIHOOD STRATEGY FOR RE-PAIRING BROKEN RAND..(U) OHIO STATE UNIV RESEARCH FOUNDATION COLUMBUS P K GOEL ET AL. JAN 86 1/1

UNCLASSIFIED OSURF-716366 AFOSR-TR-87-1176 F/G 12/3 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

UNCLAAIFI
SECURITY CLAS

AD-A186 164

DTIC FILE COPY (2)

UMENTATION PAGE

DTIC ELECTED OCT 14 1987 S D

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION A[illegible] NA	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for Public Release; Distribution Unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE NA	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) 716366	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 87-1176
6a. NAME OF PERFORMING ORGANIZATION Ohio State University Research Foundation	6b. OFFICE SYMBOL (If applicable)
6c. ADDRESS (City, State and ZIP Code) 1314 Kinnear Road Columbus, Ohio 43212	7a. NAME OF MONITORING ORGANIZATION AFOSR/NM
	7b. ADDRESS (City, State and ZIP Code) Bldg. 410 Bolling AFB, DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION AFOSR	8b. OFFICE SYMBOL (If applicable) NM
	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER AFOSR-84-0162
8c. ADDRESS (City, State and ZIP Code) Bldg. 410 Bolling AFB, DC	10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
6.1102F	2304	K3	

11. TITLE (Include Security Classification) Some Properties of Maximum Likelihood Strategy for Re-Pairing Broken Random Sample

12. PERSONAL AUTHOR(S)
Prem K. Goel and T. Ramalingam

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
interm	FROM 7/1/84 TO 6/30/86	1985; 11, 15	15

16. SUPPLEMENTARY NOTATION

17. COSATI CODES		
FIELD	GROUP	SUB. GR.
XXXXXX	XXXXXXXXX	XXXXXXXXXXXXXX

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Matching Problem, Maximum likelihood pairing, Asymptotic properties, Exchangeability, ϵ-matching

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Matching data from a bivariate population is considered when observations are available only in the form of a broken random sample. In other words, a random sample of n pairs is drawn from the population but the observed data consist of n observations on the second component and the n observations on an unknown permutation of the first component of the n pairs of data. A maximum likelihood matching strategy is revisited. The proportion of approximately correct matches (due to Yahav) is used to evaluate the performance of the pairing strategy as $n \to \infty$. The small sample behavior of this proportion is studied via a Monte-Carlo simulation in the special case of bivariate normal parent population.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ OTIC USERS ☐	UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE NUMBER (Include Area Code)	22c OFFICE SYMBOL
Brian W. Woodroofe, Major, USAF	NM (202 767-5025	AFOSR/NM

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE. UNCLASSIFIED

Some Properties of Maximum Likelihood Strategy for Re-Pairing Broken Random Sample

by Prem K. Goel and T. Ramalingam

The Ohio State University and Northern Illinois University

Technical Report #335

January, 1986

Department of Statistics
The Ohio State University

Some Properties of Maximum Likelihood Strategy for Re-Pairing Broken Random Sample[1]

By Prem K. Goel & T. Ramalingam
The Ohio State University and Northern Illinois University.

Abstract

Matching data from a bivariate population is considered when observations are available only in the form of a broken random sample. In other words, a random sample of n pairs is drawn from the population but the observed data consist of n observations on the second component and the n observations on an unknown permutation of the first component of the n pairs of data. A maximum likelihood matching strategy is revisited. The proportion of approximately correct matches (due to Yahav) is used to evaluate the performance of the pairing strategy as $n \to \infty$. The small sample behavior of this proportion is studied via a Monte-Carlo simulation in the special case of bivariate normal parent population.

Keywords and Phrases: Matching Problem, Maximum likelihood pairing, asymptotic properties, exchangeability, ε-matching,

[1]This work was supported in part by The National Science Foundation under grant number DMS-8400687 and by The Air Force Office of Scientific Research under Contract number AFOSR-84-0162.

Some Properties of Maximum Likelihood Strategy for Re-Pairing Broken Random Sample

By Prem K. Goel & T. Ramalingam
The Ohio State University and Northern Illinois University.

1. Introduction. An important tool for analyzing economic policies is the microanalytic model. Many Federal agencies use such models for the evaluation of policy proposals. When all the input-data for the model come from a single source, the quality of the model depend on, among others, how complete the information is on jointly observed variables. Often times, the input for the model consists of data from more than one Federal Agency. For instance, to make-up for 'gaps' that occur in decennial Census, the Bureau of the Census and the Internal Revenue Service provide marginal information on variables. However, joint information on these variables is not available to either of the two agencies. In such cases, Federal statisticians use file merging methodology in order to produce comprehensive data on variables of interest. A review of the origin, progress and recent developments of this methodology is given in Radner et al (1980).

An unified frame work for all such models for the file-merging methodology and statistical properties of some of them are given in Ramalingam (1985) and Goel and Ramalingam (1985). One useful model for obtaining matched pairs, introduced by DeGroot, Feder and Goel (1971) is as follows: Let $\mathbf{W}_i = (\mathbf{T}_i, \mathbf{U}_i)$, $i=1,2,...n$ be iid random vectors which are not observable as (**t, u**) pairs. Instead,it is assumed that the marginal data on **t** and **u** are available on these n individuals as follows.

File 1: $\mathbf{x}_1, \mathbf{x}_2,.....,\mathbf{x}_n$, which is an unknown permutation of the unobserved values $\mathbf{t}_1,..,\mathbf{t}_n$.

File 2: $\mathbf{u}_1, \mathbf{u}_2,...,\mathbf{u}_n$.

Thus data in File 1 is available at one agency and the data in File 2 is available at the other agency. Clearly , what is missing from the conceptually unobserved values on (**t,u**) is the pairing which, identifies the $\mathbf{t}_i$ and $\mathbf{u}_i$ that pertain to the same individual. DeGroot, Feder and Goel (1971), call the marginal observed data $\mathbf{x}_1,..,\mathbf{x}_n$; $\mathbf{u}_1,...,\mathbf{u}_n$ a *Broken Random Sample* from the population of (**T,U**).

In this paper, we shall derive some statistical properties of known stategies to merge File1 and File 2 in order to reconstruct paired data on $(\mathbf{T}_i,\mathbf{U}_i)$ for the bivariate matching problem in which both **T** and **U** are one-dimensional variables. We shall begin with some notations.

1.1 Notations. Let (T,U) have an absolutely continuous joint CDF H(t,u) and joint density h(t,u). The marginal distribution functions of T and U will be denoted by G(.) and F(.) respectively and I [.] will denote the indicator function of the event.

Let $G_n(x) = (1/n) \sum_i I[T_i \le x]$ denote the empirical CDF based on the variables $T_1,\ldots, T_n$. Similarly, $F_n(x)$ denotes the empirical CDF based on $U_1,\ldots, U_n$.

Let $R(i) = \sum_\alpha I[T_i \ge T_\alpha]$ denote the rank of T_i, $i=1,2,\ldots,n$. Similarly $S(1), \ldots, S(n)$ denote the ranks of the variables $U_1, U_2, \ldots, U_n$.

Let $\varphi = (\varphi(1),\ldots, \varphi(n))$ be a permutation of the integers $1,2,\ldots,n$. The set of all n! permutations of $1,2,\ldots,n$ will be denoted by Ψ. Let $\varphi^* = (1,2, \ldots,n)$ denote the identity permutation.

Let $\varepsilon \ge 0$. For all $i = 1,2, \ldots n$, and $\varphi \in \Psi$ define events $A_{ni}(\varphi, \varepsilon)$ and $A_{ni}(\varepsilon)$ as follows:

$$A_{ni}(\varphi, \varepsilon) \equiv \{ | U_{(\varphi(R(i)))} - U_i | \le \varepsilon \}. \tag{1.1}$$

$$A_{ni}(\varepsilon) \equiv A_{ni}(\varphi^*, \varepsilon). \tag{1.2}$$

For all $1 \le j,k \le n$, let

$$\xi_{1jk} \equiv I[U_j - U_k \ge \varepsilon] - I[T_j - T_k \ge 0], \tag{1.3}$$

$$\xi_{2jk} \equiv I[T_j - T_k \ge 0] - I[U_j - U_k \ge -\varepsilon], \tag{1.4}$$

and $\beta_1 =_d \beta_2$ denotes that the vectors β_1 and β_2 have identical distributions.

2. A Class of Matching Problems. Suppose that $h(t,u)$ has the monotone likelihood ratio (MLR) property. That is, for all reals $t_1 < t_2$ and $u_1 < u_2$, we have

$$h(t_1, u_1)\, h(t_2, u_2) \ge h(t_1, u_2)\, h(t_2, u_1). \tag{2.1}$$

If the broken random sample $x_1,\ldots, x_n, u_1,\ldots, u_n$ comes from $h(t,u)$, a typical 'matching strategy' based on permutation $\varphi \in \Psi$ can be described by pairing $x_{(i)}$ with $u_{(\varphi(i))}$. Generalizing the results of DeGroot,Feder & Goel (1971), Chew (1973) showed that if the MLR property (2.1) holds, then the strategy which maximizes the likelihood $\prod_i h(x_i, u_{\varphi(i)})$ of the parameter φ over Ψ, is to pair the ith smallest x with the ith smallest u. Note that, though the pairings in the unobserved sample (T_i, U_i), $i=1,2, \ldots, n$ are unavailable, the order-statistics of the marginal data on X and U are respectively the

same as the ordered values of T and U . Hence, we can write the mergeo file on (T,U) due to any strategy φ as

$$(T_{(i)}, U_{(\varphi(i))}) \quad ,2,\ldots,n \tag{2.2}$$

Consequently, the merged file based on the maximum likelihood pairing (MLP) mentioned above, is obtained by letting $\varphi = \varphi^*$ in (2.2).

Quality of the Merged File. Ideally, we would like to select a φ for which the file in (2.2) recovers all (T,U) pairs in the original unobserved data. It is therefore natural to consider the random variable N (φ) , the number of correct matches due to φ, as an indicator of the performance of the matching (merging) strategy φ. The optimality of φ^* subject to various criteria, e.g., maximizing the expected number of correct matches, E (N(φ)), is discussed in Ramalingam (1985).

Situations often arise where it is not crucial that, after the two files are merged, the matched pairs be exactly the same as the pairs of the original data. For example, when contingency tables analyses are contemplated for grouped data on continuous variables T and U then, in the absence of the knowledge of the pairings, we would like to reconstruct the pairs but would not worry too much as long as the u-value in any matched pair came within a pre-fixed tolerance ε (a non-negative number) of the true u-value that we would get with the ideal matching which recovers all the original pairs. This type of 'approximate matching' was first introduced by Yahav (1982) who defined ε -correct matching as follows.

Definition 1 (Yahav) . A pair $(x_{(i)}, u_{(\varphi(i))})$,[1] in the merged file (2.2), is said to be ε - correct, if $| U_{(\varphi(i))} - U_{[i]} | \le \varepsilon$, where $\varepsilon > 0$ and $U_{[i]}$ is the concomitant of $X_{(i)}$; that is the true u-value that was paired with $T_{(i)}$ in the original sample.

The number of ε -correct matches N (φ, ε), in the merged file (2.2) is given by

$$N(\varphi, \varepsilon) = \Sigma_i \, I [\, | U_{(\varphi(i))} - U_{[i]} | \le \varepsilon] \tag{2.3}$$

Note that as $\varepsilon \downarrow 0$, N(φ,ε) converges (almost surely) to N(φ) ,the number of exact matches.

The counts N(φ) and N(φ,ε) are useful indices reflecting the reliability of the merged file (2.2) resulting from φ. We shall now derive some statistical properties of N(φ^*,ε) /n. In view of the fact that Federal files often consist of a large number of records, it is clear that these asymptotic investigations are useful.

3. Asymptotic behavior of $N(\varphi^*,\varepsilon)$. We first establish a representation for $N(\varphi,\varepsilon)$ as a sum of exchangeable 0 -1 random variables. This representation will lead to an easy proof of the convergence in probability of the proportion, $N(\varphi^*,\varepsilon)/n$, of ε -correct matches due to MLP strategy. The following Lemma(See Randles and Wolfe (1979), Theorem 1.3.7, page 16) will be needed.

Lemma 1. If $\xi =_d \nu$ and K (.) is a measurable function (possibly vector valued) defined on the common support of these random vectors, then $K(\xi) =_d K(\nu)$

Proposition 1. Let $A_{ni}(\varphi,\varepsilon)$ and $N(\varphi,\varepsilon)$ be given by (1.1) and (2.3) respectively. Then, for all $\varphi \in \Psi$

$$N(\varphi,\varepsilon) = \Sigma_i I[A_{ni}(\varphi,\varepsilon)] \tag{3.1}$$

where the summands, $I[A_{ni}(\varphi,\varepsilon)]$ are exchangeable binary variables.

Proof: The order-statistic $U_{(\varphi(i))}$ and the concomitant $U_{[i]}$ of $T_{(i)}$ used in (2.3) can be written in terms of the ranks of T's and U's as follows:

$$U_{(\varphi(i))} = \Sigma_\alpha U_\alpha I[R_{2\alpha} = \varphi(i)] \tag{3.2}$$

$$U_{[i]} = \Sigma_\alpha U_\alpha I[R_{1\alpha} = i] \tag{3.3}$$

Note that $N(\varphi,\varepsilon)$ is simply a count of how many pairs in the merged file based on φ, as defined in (2.2), satisfy

$$| U_{(\varphi(i))} - U_{[i]} | \le \varepsilon. \tag{3.4}$$

If (3.4) holds for some i, then $\exists$ a j such that

$$| U_{(\varphi(i))} - U_j | \le \varepsilon. \tag{3.5}$$

In view of the continuity of (T_i,U_i), this correspondence is one-to-one. Therefore, the count $N(\varphi,\varepsilon)$ is same as the count given by

$$N(\varphi,\varepsilon) = \Sigma_\alpha I[| U_{(\phi(R(\alpha)))} - U_\alpha | \le \varepsilon] \tag{3.6}$$

Hence, (3.1) follows from (3.6) and the definition of A_{ni}, in (1.1).

In order to show the exchangeability of the summands in (3.1), note that the original samples are independent and identically distributed vectors. Therefore

$$\{ \mathbf{W}_{\alpha(1)}, \mathbf{W}_{\alpha(2)}, \ldots, \mathbf{W}_{\alpha(n)} \} =_d \{ \mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_n \} \tag{3.7}$$

where ($\alpha(1), \alpha(2), \ldots \alpha(n)$) is an arbitrary permutation of (1,2,..., n).

Define a function $\mathbf{f} \equiv (f_1, f_2, \ldots, f_n)$ from $\mathbb{R}^{2n}$ to $\mathbb{R}^n$ by

$$f_j = \begin{cases} 1 & \text{if } \Sigma_i I [b_j - b_i \geq \varepsilon] \leq \varphi (\Sigma_i I [a_j - a_i \geq 0]) \leq \Sigma_i I [b_j - b_i \geq -\varepsilon] \\ 0 & \text{otherwise ,} \end{cases} \tag{3.8}$$

for j=1,2,...,n, where $(a_1,b_1, \ldots, a_n,b_n)$ is an arbitrary point in $\mathbb{R}^{2n}$ and $\varphi \in \Psi$.It follows from (3.7) and Lemma 1 that

$$\mathbf{f} (\mathbf{W}_{\alpha(1)}, \mathbf{W}_{\alpha(2)}, \ldots, \mathbf{W}_{\alpha(n)}) =_d \mathbf{f} (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_n) . \tag{3.9}$$

Fix $j \in \{1,2,\ldots,n\}$. Then, using (3.8), we see that $f_j (\mathbf{W}_{\alpha(1)}, \mathbf{W}_{\alpha(2)}, \ldots, \mathbf{W}_{\alpha(n)})$ is the indicator function of the event

$$\Sigma_i I [U_{\alpha(j)} - U_i \geq \varepsilon] \leq \varphi (\Sigma_i I [T_{\alpha(j)} - T_i \geq 0]) \leq \Sigma_i I [U_{\alpha(j)} - U_i \geq -\varepsilon]$$

or, equivalently, in terms of the ranks $R_{11}, \ldots R_{1n}$ of the T's and the empirical CDF G_n (.) of U's,

$$G_n (U_{\alpha(j)} - \varepsilon) \leq (\varphi(R_{1\alpha(j)})/n \leq G_n (U_{\alpha(j)} + \varepsilon).$$

Since $G_n^{-1}(k/n) = U_{(k)}$, k=1,2,...,n, it follows that $f_j (\mathbf{W}_{\alpha(1)}, \mathbf{W}_{\alpha(2)}, \ldots, \mathbf{W}_{\alpha(n)})$ is 1 iff $| U_{\varphi(R1\,\alpha(j))} - U_{\alpha(j)} | \leq \varepsilon$. Consequently,

$$f_j (\mathbf{W}_{\alpha(1)}, \mathbf{W}_{\alpha(2)}, \ldots, \mathbf{W}_{\alpha(n)}) = I [A_{n\alpha(j)} (\varphi,\varepsilon)]. \tag{3.10a}$$

Similarly,

$$f_j (\mathbf{W}_1, \ldots, \mathbf{W}_n) = I [A_{nj} (\varphi,\varepsilon)]. \tag{3.10b}$$

The exchangeability of the summands in (3.1) follows from (3.9), (3.10a) and (3.10b).

We shall now review some results concerning E [N (ε)/n] , due to Yahav (1982), where $N(\varepsilon) \equiv N (\varphi *, \varepsilon)$. Assuming that the distribution of T and U satisfies:

the conditional distribution of U given that T=t is (univariate) normal with mean t and variance 1 ,

Yahav (1982) derived the limiting value of $\mu_n (\varepsilon)=E [N (\varepsilon)/n]$ as $n \to \infty$ by using the representation (2.3) in which the summands are functions of the order-statistics of

$U_1,\ldots, U_n$ and the concomitants of the order-statistics of $T_1, \ldots, T_n$. His proof relied on an approximation theorem, about the order-statistics for the above model, given in Bickel and Yahav(1977). Furthermore, he also reported the findings of a Monte-Carlo study for $\mu_n(\varepsilon)$ in a particular case of his model, namely, T and U are bivariate normal with correlation ρ.

We now establish the large-sample behavior of $N(\varepsilon)/n$ in case of samples from an *arbitrary population*. The properties of its expected value follow as a consequence. In section 4, we indicate how Yahav's simulation study of the small-sample properties of $\mu_n(\varepsilon)$ can be improved upon. We shall then present the results of our Monte-Carlo study of $\mu_n(\varepsilon)$ when n is small.

Theorem 1. For broken random samples from an absolutely continuous distribution,

$$N(\varepsilon)/n \rightarrow_{pr} \mu(\varepsilon), \quad \text{as } n \rightarrow \infty \tag{3.11}$$

where

$$\mu(\varepsilon) = P[\, G(U-\varepsilon) \leq F(T) \leq G(U+\varepsilon)\,] \tag{3.12}$$

Proof: Let $L_n = N(\varepsilon)/n$. Using the definitions of $A_{ni}(\varepsilon)$ in (1.2) and the representation (3.1) for $N(\varepsilon)$ as a sum of exchangeable binary variables we obtain

$$N(\varepsilon) = \Sigma_i\, I\,[A_{ni}(\varepsilon)]\,. \tag{3.13}$$

It follows that

$$E(L_n) = n\,P(A_{n1}(\varepsilon))/n = P(A_{n1}(\varepsilon))\,. \tag{3.14}$$

Note that

$$E(L_n^2) = n^{-2}\,[\,E(N(\varepsilon))^{(2)} + E(N(\varepsilon))] \tag{3.15}$$

where $E(N(\varepsilon))^{(2)}$ is the second factorial moment of $N(\varepsilon)$. Using the representation (3.13), we get

$$E(L_n^2) = n^{-2}\,[n^{(2)}\,P\{A_{n1}(\varepsilon)\,A_{n2}(\varepsilon)\} + n\,P(A_{n1}(\varepsilon))].$$

For $\alpha = 1,2,\ldots,n$, and $j=1,2$ let

$$\nu_{j\alpha} = \Sigma_i\ \xi_{j\alpha i} \tag{3.16}$$

where the sequences $\{\xi_{1\alpha i}\}$ and $\{\xi_{2\alpha i}\}$ are defined in (1.3) and (1.4). It follows that

$$A_{n1}(\varepsilon) = (\nu_{11}/n \leq 0, \nu_{21}/n \leq 0) \tag{3.17}$$

and

$$A_{n1}(\varepsilon)\,A_{n2}(\varepsilon) = \cap_i \cap_j\ (\nu_{ij}/n \leq 0)\,. \tag{3.18}$$

Note that, given $\mathbf{W}_1 = (t_1, u_1)$, the infinite sequence $\xi_{112}, \xi_{113}, \ldots$ is exchangeable. Hence, by the Strong Law of Large Numbers for exchangeable random variables (see Chow and Teicher, 1978, p.223),

$$\nu_{11}/n \rightarrow E(\xi_{112} \mid \mathbf{W}_1) \text{ a.s.} \quad \text{as } n \rightarrow \infty,$$

where the conditional expectation is equal to $\{G(u_1 - \varepsilon) - F(t_1)\}$. It follows that

$$\nu_{11}/n \rightarrow G(U_1 - \varepsilon) - F(T_1), \text{ a.s.} \quad \text{as } n \rightarrow \infty. \tag{3.19}$$

We can show by similar arguments that

$$\nu_{1\alpha}/n \rightarrow G(U_\alpha - \varepsilon) - F(T_\alpha), \text{ a.s.} \tag{3.20}$$

and

$$\nu_{2\alpha}/n \rightarrow F(T_\alpha) - G(U_\alpha + \varepsilon), \text{ a.s.}, \tag{3.21}$$

where $\alpha = 1,2$. Using the fact (see Serfling, 1980, p.52) that a sequence of vectors converges almost surely to a given vector iff the componentwise sequences converge almost surely to the appropriate components of the limit, we get from (3.20) and (3.21)

$$\begin{bmatrix} \nu_{11}/n \\ \nu_{21}/n \\ \nu_{12}/n \\ \nu_{22}/n \end{bmatrix} \rightarrow \begin{bmatrix} G(U_1 - \varepsilon) - F(T_1) \\ F(T_1) - G(U_1 + \varepsilon) \\ G(U_2 - \varepsilon) - F(T_2) \\ F(T_2) - G(U_2 + \varepsilon) \end{bmatrix} \quad \text{a.s.} \tag{3.22}$$

It follows from (3.17), (3.18), (3.22) and the independence of $\mathbf{W}_1$ and $\mathbf{W}_2$ that

$$P(A_{n1}(\varepsilon)) \rightarrow \mu(\varepsilon) \tag{3.23}$$

and

$$P(A_{n1}(\varepsilon) A_{n2}(\varepsilon)) \rightarrow \mu^2(\varepsilon). \tag{3.24}$$

Therefore (3.14), (3.15) and (3.23), (3.24) imply that as $n \rightarrow \infty$,

$$E(L_n) \rightarrow \mu(\varepsilon), \tag{3.25}$$

and

$$\mathrm{Var}(L_n) \rightarrow 0. \tag{3.26}$$

It is well known that (3.25) and (3.26) imply the convergence in probability as in (3.11).

The following corollary generalizes Yahav's result concerning $\mu_n(\varepsilon)$, the first moment of $N(\varepsilon)/n$.

Corollary1 . For p>0, L_p

(i) $N(\varepsilon)/n \rightarrow \mu(\varepsilon)$ as $n \rightarrow \infty$

(i i) $E[(N(\varepsilon)/n)^p] \rightarrow [\mu(\varepsilon)]^p$ as $n \rightarrow \infty$.

Proof: The number of ε -correct matches can atmost be n, the number of pairs in the unobserved bivariate-data.Therefore $0 \leq N(\varepsilon)/n \leq 1$, for all n =2,3,... and $\{N(\varepsilon)/n\}$ is a uniformly bounded sequence of random variables. It is well known that convergence in probability and L_p-convergence are equivalent for such sequences. Hence (i) follows easily from Theorem 2. Now, (ii) readily follows from (i) because $(\mu(\varepsilon)^p)$ is finite for p>0.

Note that in our results, no assumption about the conditional distribution of U given T has been made as was the case with Yahav's results.

4. Small Sample behavior of N(ε). Yahav used simulated samples from a bivariate - normal population with mean vector **0** and covariance matrix

$$\Sigma = (1-\rho^2)^{-1} \begin{bmatrix} \rho^2 & \rho^2 \\ \rho^2 & 1 \end{bmatrix} , \qquad (4.1)$$

to study small sample properties of $\mu_n(\varepsilon)$. It is important to note that in (4.1), the variances of T and U are functions of their correlation, ρ. This is so, because Yahav's model requires that the conditonal distribution of U given T=t be normal with mean t and variance 1. The limiting value of $\mu_n(\varepsilon)$ for his particular model is given by:

$$\mu(\varepsilon) = \int_{\mathfrak{R}} \{\Phi(x\, a(\rho) + \varepsilon/\rho) - \Phi(x\, a(\rho) - \varepsilon/\rho)\}\, d\Phi(x) , \qquad (4.2)$$

where $a(\rho) = [(1-\rho)/(1+\rho)]^{1/2}$.

Yahav computed $\mu(\varepsilon)$ by numercal integration for ε = 0.01, 0.05, 0.1 & 0.3. However, it can be shown that (4.2) simplifies to

$$\mu(\varepsilon) = 1 - 2\,\Phi\,[-((1+\rho)/2)^{-1/2})\,\varepsilon/\rho\,]. \qquad (4.3)$$

Yahav also provided Monte-Carlo estimates of $\mu_n(\varepsilon)$, for n = 10, 20,50 and 100 using the simulated data on T and U. Table 4.1 is a typical example of one of his results.

Table 4.1 Expected Average Number of ε-correct Matchings, ε = .01 [Yahav(1982)]

ρ	$\mu_{10}(\varepsilon)$	$\mu_{20}(\varepsilon)$	$\mu_{50}(\varepsilon)$	$\mu_{100}(\varepsilon)$
.001	.5864	.5326	.5275	.5227
.01	.1984	.1648	.1271	.1152
.10	.1512	.1058	.0760	.0591
.30	.1084	.0686	.0389	.0214
.50	.1020	.0582	.0272	.0138
.70	.0960	.0614	.0262	.0105
.90	.0972	.0540	.0206	.0086
.95	.0976	.0496	.0214	.0083
.99	.0960	.0484	.0213	.0080

It is clear from Table 4.1 and equation (4.3) that $\mu_n(\varepsilon)$ and $\mu(\varepsilon)$ decrease as ρ ranges from 0.001 to 0.99. In fact, (4.3) implies that $\mu(\varepsilon)=1-2\Phi(-\varepsilon)$ for $\rho=1.0$ and $\mu(\varepsilon) = 1.0$ for $\rho=0$, which goes against the intuition. One expects that for an optimal strategy, such as φ^*, $\mu_n(\varepsilon)$ as well as $\mu(\varepsilon)$ must be monotone increasing in ρ. The problem here is not with the MLP φ^*, but with the covariance matrix Σ, defined by (4.1), used in Yahav's model. Because, as ρ changes its value, so do the marginal variances of T and U. In fact, as $\rho \rightarrow 1$, the marginal variances $\rightarrow \infty$. To rectify this problem, we have assumed a bivariate normal model for T and U with means zero, variances one and the correlation ρ.

For each combination of four values of n, namely 10, 20, 50 and 100, and twelve values of ρ, namely 0.00, 0.10 (0.10) 0.90, 0.95, 0.99; 1000 sample were generated from the bivariate normal population using the IMSL Library routines. These data were used to obtain Monte-Carlo estimates of $\mu_n(\varepsilon)$, where ε was given the values 0.01, 0.05, 0.1, 0.3, 0.5, 0.75, 1.0.

It is easy to show that, for the above model

$$\mu(\varepsilon) = P(|Z| \leq \varepsilon (2(1-\rho))^{-1/2}), \tag{4.4}$$

where Z is a standard normal random variable. It is clear from (4.4) that $\mu(\varepsilon)$ is a monotone increasing function of ρ. Using standard-normal CDF tables, $\mu(\varepsilon)$ in (4.4) was computed for each combination of the twelve values of ρ and the seven values of ε mentioned above. The estimated values of $\mu_n(\varepsilon)$ and the limiting value $\mu(\varepsilon)$ are given in Tables A.1-A.7 in the Appendix.

Note that, as expected, $\mu_n(\varepsilon)$ and $\mu(\varepsilon)$ in Tables A.1-A.7 are monotone increasing functions of ρ for each fixed ε. Furthermore, the quality of the merged file is quite good if we want to reconstruct contingency tables with intervals of size .5 σ or more and the correlation $\rho \geq 0.5$.

APPENDIX

Table A.1 Expected Average number of ε-correct Matchings, $\varepsilon = 0.01$

ρ	$\mu_{10}(\varepsilon)$	$\mu_{20}(\varepsilon)$	$\mu_{50}(\varepsilon)$	$\mu_{100}(\varepsilon)$	$\mu(\varepsilon)$
0.00	0.106	0.054	0.025	0.015	0.008
0.10	0.113	0.059	0.028	0.017	0.008
0.20	0.127	0.068	0.031	0.018	0.008
0.30	0.138	0.075	0.034	0.020	0.008
0.40	0.155	0.083	0.038	0.023	0.008
0.50	0.174	0.095	0.044	0.026	0.008
0.60	0.199	0.109	0.061	0.036	0.008
0.70	0.231	0.129	0.061	0.036	0.008
0.80	0.279	0.162	0.077	0.046	0.016
0.90	0.374	0.222	0.109	0.067	0.016
0.95	0.476	0.296	0.151	0.094	0.024
0.99	0.700	0.521	0.299	0.191	0.056

Table A.2 Expected Average number of ε-correct Matchings, $\varepsilon = 0.05$

ρ	$\mu_{10}(\varepsilon)$	$\mu_{20}(\varepsilon)$	$\mu_{50}(\varepsilon)$	$\mu_{100}(\varepsilon)$	$\mu(\varepsilon)$
0.00	0.127	0.076	0.047	0.037	0.032
0.10	0.134	0.082	0.051	0.040	0.032
0.20	0.149	0.093	0.056	0.043	0.032
0.30	0.161	0.099	0.061	0.047	0.032
0.40	0.180	0.109	0.066	0.052	0.040
0.50	0.201	0.124	0.074	0.057	0.040
0.60	0.228	0.141	0.085	0.065	0.048
0.70	0.262	0.166	0.101	0.076	0.048
0.80	0.317	0.205	0.124	0.094	0.064
0.90	0.420	0.280	0.174	0.135	0.088
0.95	0.529	0.368	0.237	0.186	0.127
0.99	0.769	0.631	0.459	0.377	0.274

Table A.3 Expected Average number of ε-correct Matchings, $\varepsilon = 0.1$

ρ	$\mu_{10}(\varepsilon)$	$\mu_{20}(\varepsilon)$	$\mu_{50}(\varepsilon)$	$\mu_{100}(\varepsilon)$	$\mu(\varepsilon)$
0.00	0.154	0.102	0.075	0.065	0.056
0.10	0.160	0.110	0.080	0.069	0.056
0.20	0.177	0.121	0.087	0.074	0.064
0.30	0.189	0.130	0.093	0.080	0.064
0.40	0.210	0.143	0.101	0.088	0.072
0.50	0.234	0.161	0.112	0.096	0.080
0.60	0.264	0.181	0.127	0.108	0.088
0.70	0.302	0.210	0.149	0.126	0.103
0.80	0.363	0.258	0.182	0.154	0.127
0.90	0.477	0.347	0.254	0.218	0.174
0.95	0.594	0.452	0.342	0.299	0.251
0.99	0.839	0.744	0.630	0.580	0.522

Table A.4 Expected Average number of ε-correct Matchings, $\varepsilon = 0.3$

ρ	$\mu_{10}(\varepsilon)$	$\mu_{20}(\varepsilon)$	$\mu_{50}(\varepsilon)$	$\mu_{100}(\varepsilon)$	$\mu(\varepsilon)$
0.00	0.255	0.208	0.184	0.175	0.166
0.10	0.265	0.223	0.195	0.186	0.174
0.20	0.284	0.237	0.207	0.197	0.190
0.30	0.305	0.253	0.221	0.211	0.197
0.40	0.334	0.275	0.240	0.229	0.213
0.50	0.363	0.304	0.263	0.250	0.236
0.60	0.401	0.336	0.293	0.278	0.266
0.70	0.455	0.382	0.337	0.320	0.303
0.80	0.532	0.457	0.403	0.386	0.362
0.90	0.670	0.593	0.540	0.519	0.497
0.95	0.802	0.733	0.689	0.674	0.658
0.99	0.978	0.968	0.961	0.961	0.966

Table A.5 Expected Average number of ε-correct Matchings, $\varepsilon = 0.5$

ρ	$\mu_{10}(\varepsilon)$	$\mu_{20}(\varepsilon)$	$\mu_{50}(\varepsilon)$	$\mu_{100}(\varepsilon)$	$\mu(\varepsilon)$
0.00	0.353	0.311	0.290	0.281	0.274
0.10	0.363	0.330	0.306	0.298	0.289
0.20	0.390	0.348	0.325	0.315	0.311
0.30	0.417	0.371	0.344	0.336	0.326
0.40	0.452	0.400	0.373	0.362	0.354
0.50	0.485	0.437	0.404	0.393	0.383
0.60	0.528	0.478	0.446	0.435	0.425
0.70	0.591	0.536	0.506	0.495	0.484
0.80	0.675	0.628	0.594	0.584	0.570
0.90	0.811	0.773	0.752	0.744	0.737
0.95	0.917	0.896	0.888	0.885	0.886
0.99	0.998	0.999	0.999	0.999	1.000

Table A.6 Expected Average number of ε-correct Matchings, ε = 0.75

ρ	$\mu_{10}(\varepsilon)$	$\mu_{20}(\varepsilon)$	$\mu_{50}(\varepsilon)$	$\mu_{100}(\varepsilon)$	$\mu(\varepsilon)$
0.00	0.468	0.433	0.416	0.409	0.404
0.10	0.488	0.454	0.437	0.429	0.425
0.20	0.514	0.477	0.461	0.453	0.445
0.30	0.539	0.505	0.487	0.480	0.471
0.40	0.582	0.542	0.522	0.514	0.503
0.50	0.621	0.586	0.560	0.555	0.547
0.60	0.662	0.633	0.613	0.606	0.59
0.70	0.727	0.694	0.679	0.673	0.668
0.80	0.810	0.786	0.772	0.768	0.766
0.90	0.919	0.908	0.906	0.904	0.907
0.95	0.979	0.976	0.978	0.979	0.982
0.99	1.000	1.000	1.000	1.000	1.000

Table A.7 Expected Average number of ε-correct Matchings, ε = 1.0

ρ	$\mu_{10}(\varepsilon)$	$\mu_{20}(\varepsilon)$	$\mu_{50}(\varepsilon)$	$\mu_{100}(\varepsilon)$	$\mu(\varepsilon)$
0.00	0.570	0.545	0.531	0.524	0.522
0.10	0.593	0.566	0.555	0.549	0.547
0.20	0.621	0.595	0.581	0.576	0.570
0.30	0.646	0.622	0.611	0.605	0.605
0.40	0.690	0.664	0.650	0.644	0.627
0.50	0.729	0.707	0.691	0.688	0.683
0.60	0.772	0.753	0.744	0.741	0.737
0.70	0.830	0.812	0.807	0.805	0.803
0.80	0.898	0.889	0.887	0.885	0.886
0.90	0.970	0.970	0.972	0.972	0.975
0.95	0.996	0.996	0.997	0.997	0.998
0.99	1.000	1.000	1.000	1.000	1.000

REFERENCES

Bickel, P.J and Yahav, J.A. (1977) On Selecting a Subset of Good Populations, in *Statistical Decision Theory and Related Topics* II, (Eds.) S. S. Gupta and D.S. Moore, Academic Press, New York.

Chow, Y.S and Teicher, H.(1978) *Probability Theory,* Springer-Verlag, New York.

Chew, M.C (1973), On Pairing Observations from a Distribution with Monotone Likelihood Ratio, *Annals of Statistics,* **1** , 433-445.

DeGroot, M.H, Feder, P.I., Goel, P.K. (1971), Matchmaking, *Annals of Mathematical Statistics*, **42**, 578-593.

Goel, P.K. & Ramalingam, T. (1985) The Matching Methodology:Some Statistical Properties,Technical Report # **333**,Department of Statistics,Ohio State University

Radner, D.B. et. al (1980), *Report on Exact and Statistical Matching Techniques*, Statistical Policy Working Paper No. **5**, Office of Federal Statistical Policy and Standards, U.S. Dept. of Commerce.

Ramalingam, T (1985): *Statistical Properties of the File-merging Methodology,* Ph.D. Thesis, Purdue University

Randles, R.H and Wolfe, D.A (1979), *Introduction to the Theory of Nonparametric Statistics*, John Wiley, New York.

Serfling, R.J (1980), *Approximation Theorems of Mathematical Statistics,* John Wiley, New York.

Yahav, J. A. (1982), On Matchmaking, in *Statistical Decision Theory and Related Topics* III, vol 2.,(Eds.) S.S. Gupta and J.O. Berger,497-504 Academic Press, New York.

END

12-87

DTIC